



השוואת ביצועי מודלים גדולים של שפה בהכשרה לשיקום וסטיבולרי ChatGPT, Google Gemini ואנשי מקצוע קליניים

יעל ארבל P.T., MHA^{1,2}, יואב גימון P.T., PhD^{2,3}, ליאורה שמואלי PhD¹

¹ המחלקה לניהול מערכות בריאות, אוניברסיטת בר-אילן, רמת גן, ² שירותי בריאות כללית – מכון פיזיותרפיה קריית-אונן, מחוז דן-פתח תקווה ³ החוג לפיזיותרפיה, אוניברסיטת חיפה מחלקת אף-אוזן-גרונ וניתוחי ראש-צוואר, מרכז רפואי שיבא, תל השומר

רקע

- שיקום וסטיבולרי דורש ידע קליני מדויק ושיקול דעת מקצועי.
- שילוב טכנולוגיות AI כגון LLMs (ChatGPT ו-Google Gemini) מציעים פוטנציאל בלמידה ובהדרכה.
- יישומם בשיקום וסטיבולרי טרם נבחן באופן שיטתי.
- מחקר זה מעריך את הדיוק והאמינות של המידע המסופק על ידי המודלים בתחום השיקום והוסטיבולרי.

מטרה

מטרת המחקר הייתה להשוות את ביצועי ChatGPT ו-Google Gemini מול פיזיותרפיסטים מנוסים וסטודנטים במבחן ידע וסטיבולרי, ולבחון את איכות ודיוק ההסברים שנותן המודל על ידי מומחים אוטונומולוגים.

שיטה

שלושים פיזיותרפיסטים מנוסים ושלושים סטודנטים השלימו את מבחן הידע הווסטיבולרי (VKT) בן 20 השאלות, המחולקות לשלוש קטגוריות (ידע קליני, פרקטיקה קלינית בסיסית וחשיבה קלינית), ChatGPT ו-Gemini התבקשו להשיב על אותן שאלות, ושלושה מומחים אוטונומולוגים העריכו את התשובות לפי סולם בן ארבע רמות.

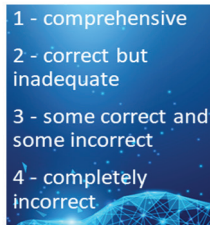
Validity VKT questionnaire

ChatGPT and Google Gemini prompted VKT

Grading ChatGPT response

Statistical analyses and results

Discussion and conclusion



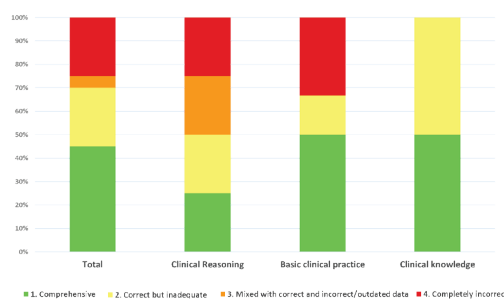
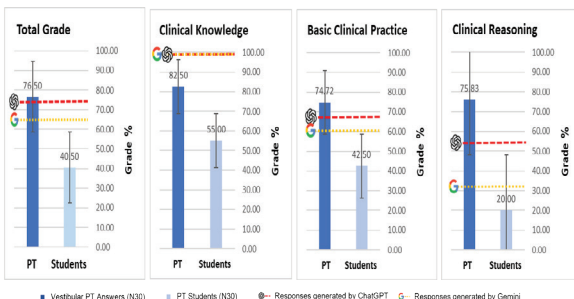
ביצועי ChatGPT, Gemini, פיזיותרפיסטים וסטודנטים במבחן ה-VKT

Google Gemini ו-ChatGPT השיגו שיעורי דיוק של 70% ו-60% בהתאמה, הצטיינו בידע הקליני (100%) אך הפגינו ביצועים נמוכים בחשיבה קלינית (25%-50%). בהשוואה, פיזיותרפיסטים השיגו ציון ממוצע של 76.5% והסטודנטים 40.5% במבחן ה-VKT.

הערכת מומחים לדיוק ההסברים של ChatGPT-3.5

תוצאה הכללית: 45% מתשובות של ChatGPT דורגו כמקיפות (Comprehensive) ו-25% כשגויות לחלוטין (Completely incorrect). בקטגוריות ה"ידע הקליני" (Clinical Knowledge) והקליניקה הבסיסית (Basic Clinical Practice) 50% מהתשובות היו מקיפות, אך רק 25% מהתשובות בקטגוריית החשיבה הקלינית (Clinical Reasoning) היו מקיפות.

תוצאות



ChatGPT ו-Google Gemini מצטיינים במענה לשאלות בתחום הידע הקליני אך מוגבלים בחשיבה קלינית, כאשר ChatGPT עלה בביצועיו על Gemini. שימוש מושכל מחייב ניסוח שאלות מדויקות ואימות מקצועי, בשל הסתמכותם על מידע מקוון. שילוב מומחיות קלינית והנחיות מקצועיות עם בינה מלאכותית עשוי לשפר קבלת החלטות ולהפחית סיכונים. גרסאות עתידיות צפויות להציג שיפור בדיוק.

מסקנות